DOCUMENT RESUME

ED 290 156 CS 211 017

AUTHOR Hawkins, Mary Louise; Marshall, Jon C.
TITLE Evaluating a Writing Program, Grades 3-12.
INSTITUTION Ferguson-Florissant School District, Ferguson, Mo.
PUB DATE 81
NOTE 40p.; Contents of this publication were developed under a grant from Title II Basic Skills, Department of Education.
AVAILABLE FROM Ferguson-Florissant Writers Project, c/o Mary Ann Croddy, Director, 1005 Waterford Drive, Florissant, MO 63033 ($4.50).
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Elementary Secondary Education; Evaluation Methods; Holistic Evaluation; Program Descriptions; *Program Evaluation; *Teaching Methods; *Writing Evaluation; *Writing Improvement; *Writing Instruction; Writing Research; Writing Skills
IDENTIFIERS Bay Area Writing Project

ABSTRACT

Based on a writing program evaluation project that took place in the Ferguson-Florissant School District in suburban St. Louis, Missouri, this booklet provides administrators and policymakers with a plan for evaluating their own writing programs and also provides results of the Ferguson-Florissant evaluation. The first chapter contains introductory material on the importance of writing as a basic skill for young people, and describes the background and development of the writing program, which was based on the Bay Area Writing Project. The assessment project's methodology is spelled out in the second chapter, while the third chapter focuses on the research design and analysis. In the fourth chapter, the sampling procedures are described, while issues and procedures for data collection are emphasized in the fifth chapter (particularly holistic scoring methods). The sixth chapter explains how the data--six essays by each participating student--were prepared for scoring; the seventh chapter includes details of how the data were actually scored. Data analysis methods and the results of the Ferguson-Florissant study are contained in the eighth chapter, which reveals that students of teachers trained in the Bay Area Writing Project Model obtained the highest mean scores. The final chapter contains concluding remarks about the program's success, including seven principles about writing program evaluation. (SKC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

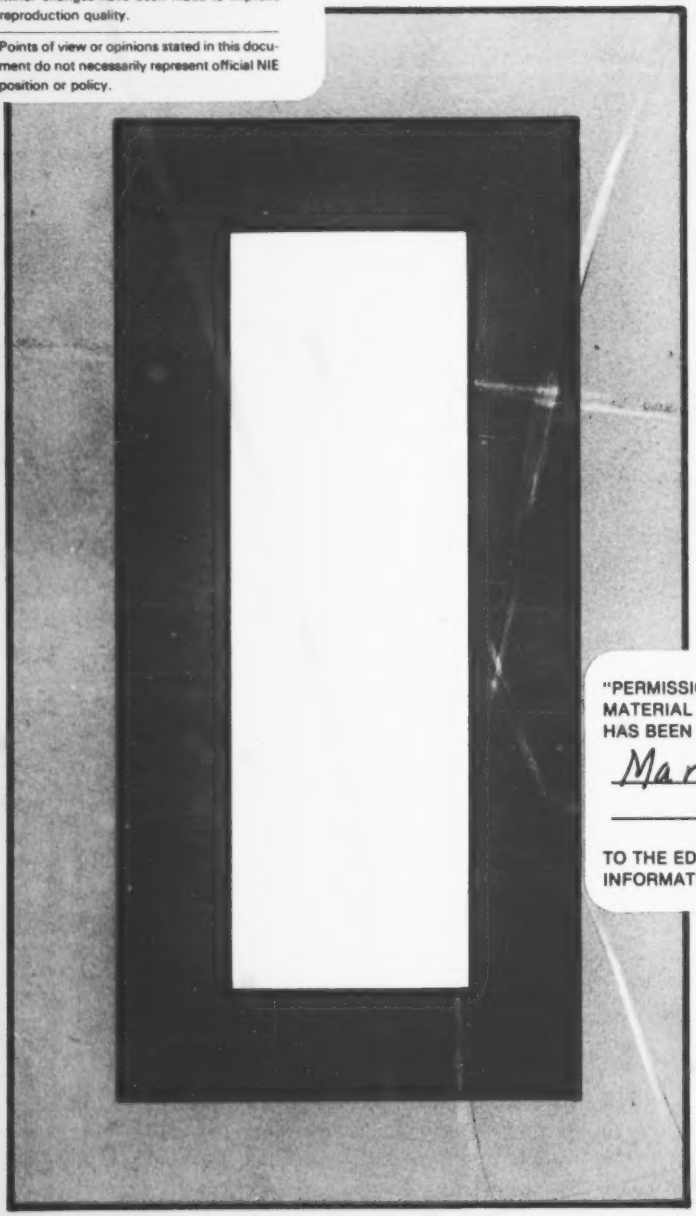
☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED290156

Evaluating A Writing Program



Grades 3 - 12

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Mary Ann Craddy

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS211017

Evaluating Writing

Grades 3 - 12

Prepared By:

Dr. Mary Louise Hawkins
Ferguson-Florissant School District
1005 Waterford Drive
Florissant, MO 63033

Dr. Jon C. Marshall
University of Missouri-St. Louis
St. Louis, Missouri

© 1981

The contents of this publication were developed under a grant from Title II Basic Skills Department of Education. The contents do not necessarily represent the policies of those agencies nor endorsement by the Federal Government.

Foreword

An automatic add-on to the familiar "Johnny can't read" refrain is "Johnny can't write." And more recently has been added the further admonition, "And his teacher can't write, either."

These refrains that haunt educators are now sung more loudly and flung wider than teachers ever imagined. Indeed, in the 80's, they carry an aura of truism almost above examination or thoughtful consideration. The sheer weight of these criticisms demands that educational organizations seriously review their own efforts to teach writing to determine for themselves the state of order of their own houses.

The Ferguson-Florissant School District, in suburban St. Louis County, Missouri, targeted writing as a priority skill area for improvement. The Bay Area Writing Project was identified as the most promising new approach. Through a Title II Basic Skills grant, the requisite teacher training and support services were implemented.

Even though all available information suggests that the instructional theory of the Bay Area program is sound, little control group research was available to document success. In addition, a

program strategy that is effective on the West Coast may or may not achieve the same result in the Midwest. Thus, an essential ingredient of the implementation plan was research to determine the impact of the process on student writing in the Ferguson-Florissant School District. The evaluation would also yield baseline data to determine the degree of student growth in writing skill that typically could be expected. This report is a description of the program evaluation process that was used and the results of the evaluation.

Project direction and maintenance was provided by Dr. Mary Louise Hawkins and Donna O'Kief, Ferguson-Florissant School District. Data management and analysis was performed by Dr. Jon C. Marshall, and the scoring of writing samples was coordinated by Jane Zeni Flynn, both from the University of Missouri-St. Louis. The teacher training program was led by Dr. Ruie Pritchard, North Carolina State University, and received its inspiration from Dr. Ben Nelms, Director of the Missouri Writers Project, University of Missouri-Columbia. The successful writing instruction that is now ongoing in the Ferguson-Florissant School District is a tribute to those teachers who were willing to examine their methods and to attempt new teaching strategies. And, the achievement demonstrated is a tribute to many students who can now legitimately say "Maybe we couldn't, but now we're learning."

Chapter I

Introduction

Writing has come to the foreground as a critical basic skill for young people. As a consequence, writing has received increasing amounts of attention by educational researchers who have crystallized through their findings a more definite route to successful writing instruction.

At the source of the newfound direction for writing instruction are two tenets:

1. **Writing is an act of integration.** It is a synergistic act; a summative expression of personal experience, skill, perspective, sense of purpose, personal style, and form. It seems appropriate, then, that writing be taught holistically and distinctly rather than as a by-product of English mechanics, such as punctuation, grammar or spelling.

2. **Writing occurs in four distinct stages, all of which are priority areas for instruction.** Thus, writing instruction should focus on the processes of writing rather than just the final product.

Embracing these two tenets, writing educators wishing to

vitalize their programs have developed new teaching strategies and implemented teacher inservice programs. The emergent strategies include sentence modeling, sentence combining, framed paragraphs, holistic grading, peer response, and free writing. The associated inservice models include, as consistent fare, research review, teacher presentations, collegial response groups and personal writing.

During the summer of 1980, 15 teachers, 12 with teaching assignments, attended a four week writer's workshop patterned after those developed by the Bay Area Writers Project and subsequently adopted by many other groups loosely connected through the National Diffusion Network or through various state agencies. The teachers represented seven different schools in levels spanning grades four through twelve.

During the fall of 1980, the trained teachers began to implement their new skills and knowledge in their own classrooms. In addition, they were charged with the responsibility of enlisting the participation of other teachers in their own schools. To this end, they made both formal faculty presentations and informal contacts with colleagues. This initial group met on a monthly basis to continue their personal writing, share experiences and interact with outside consultants. Their published writings were disseminated throughout the district. Drop-in writing centers were established in two schools. Demonstration teaching was provided upon request by the Basic Skills Title II staff. These "writing" teachers had both special training and support services that were not available to other district teachers.

The writing program existing in the other 21 district schools in grades 3-12 could be described as eclectic. At the elementary grades, writing instruction was left to the discretion of the teacher. Some teachers attended to the task more systematically than

others. At the secondary level, students were expected to produce reasonably literate paragraphs by the end of the ninth grade. Experiences at the high school level varied depending upon the school and the student. At one school, students had the opportunity to select special quarter courses in writing, beginning with paragraph skills and progressing to multiparagraph essays and research papers. At the other two district high schools, writing instruction was integrated into more typical yearlong English courses. No systematic writing assessment had been performed in the district, although language skills were tested at regular grade intervals using the **Comprehensive Test of Basic Skills** and **Test of Basic Competencies,** with the **Scholastic Aptitude Test** and **American College Test** administered to eligible students upon request. These conditions provided a prime opportunity to collect data that could be analyzed to provide information on the effectiveness of the new program.

Group	Grade	Subject	Number
Elementary	4	---	2
	6	---	1
Middle	7,8,9	English	4
High School	10,11,12	English	4
		Social Studies	1

Table 1 – Distribution of Trained Teachers

Chapter II

Methodology

Although the intent to assess students' writing skill is a worthy one, it sometimes seems to be, instead, a lofty mission impossible. The problems inherent in such an undertaking are easily enumerated, but not so easily resolved.

Foremost is the issue of how student writing can be assessed in any but a subjective way. Subjective grading may be adequate for instructional purposes, but often it falls far short of yielding the quantifiable results necessary for objective analysis. On the other hand, the more content specific objective methodologies such as multiple choice measures yield appropriately manipulatable data but do not require students to write.

Another aspect of the assessment methodology issue is the belief that any assessment procedure should deal with the whole of a piece of writing rather than measure, count, or weigh its specific parts.

In the early literature (Marshall and Hales, 1971), it was noted that reliable assessment could be accomplished through the development and utilization of detailed keys to "point-score" specific content. A sorting method (Marshall and Hales, 1971) also was recommended for classroom use in which the

reader would scan through several papers to develop a set of general criteria for sorting the papers into groups. It was pointed out that this latter method seemed the more valid since it placed primary weight on the presence of higher order processes of organization, analysis and synthesis in the product rather than factual knowledge; however, problems in scoring consistency and reliability were noted.

More recently, a third method of writing assessment called "holistic scoring" (Myers, 1980) has emerged that seems to have incorporated the "point-score" specifications (now called a "rubric") with the philosophy of the sorting method. This procedure provides a quick, reliable method of assessing writing skill through a general impression of a paper's characteristics.

Holistic scoring has gained stature in recent years. The National Assessment of Educational Progress has used holistic scoring to measure the writing competency of students in national samples. The National Council of Teachers of English also has lent its weight to the procedure, in good measure through the publications of *A Procedure for Writing Assessment and Holistic Scoring* (Miles, 1980).

Behind the practice of holistic scoring are some foundational assumptions. According to the Educational Testing Service these include:

1. the unit to be measured is the whole of a piece of writing, not its individual parts

2. good writing, although difficult to define, is easily recognizable

3. although not all educators would agree on the weight

to assign a particular trait, they would rank papers in a similar way when judging the work as a whole.

These theoretical assumptions point the way toward a method of holistic scoring. One would read a sample piece of writing, giving it a single impression score. Papers then would be judged against some reader's internal standards. To ensure that multiple readers would use the same internal standards, the scoring range and prototypes for each score value would be predetermined from the set of papers to be assessed, thus promoting consensus among readers and establishing a more standardized set of criteria for scoring.

The use of the holistic method of assessment provides meaningful data to assess both the general state of writing across a district and to measure the impact of any adjustments being made in the writing instructional program.

Another issue of concern in a writing assessment is determining at which point consideration for students' instructional opportunities should supersede the sometimes strict demands of successful research. Or, at least, in what ways can an easy peace between the two be struck.

One of the most powerful research procedures is to compare pretest and posttest scores from experimental groups with pretest and posttest scores of control groups. This design is enhanced when students and teachers are randomly assigned to be either in the experimental group or the control group. This methodology implies that program services be deliberately withheld from some students and teachers, and that students' classroom assignments be determined by research requirements rather than by instructional concerns. Imposing such a procedure on students, teachers and administrators in most school districts is not only

difficult but probably not even desirable.

Even though the ideals of research and the realities of instruction are often incompatible, some basic parameters in the evaluative process could be met in the evaluation undertaken. Since teachers volunteered for training, an important criteria for staff development programs, these specially trained teachers and their students formed a natural treatment group. Both pretests and posttests could be administered to assess whether or not change occurred during the treatment period. Also, a control group could be used since some teachers and students did not receive services. These control groups would provide the data to ascertain whether any change that might be documented could be accounted for by history effects, the instrumentation used, maturation or other factors not relevant to the new instructional program. Because random assignment of students to the treatment group was not possible, an ex post facto pretest-posttest control group design was used where students were placed in either treatment or control groups based on whether or not they were taught by trained teachers.

Another issue that must be resolved in evaluating student writing is the tremendous number of administrative tasks and concerns that must be anticipated lest the whole process be drowned in a sea of student papers. Any English teacher who ends a teaching day with 150 student papers to grade recognizes the problem. Imagine, though, a day when the student paper pile mounts to over 17,000. Resources need to be identified and procedures carefully detailed to meet the day that eventually comes.

These methodological issues, once recognized, can be managed as simply part of the evaluation process.

Chapter III

Research Design

An ex post facto pretest-posttest control group design was selected that would allow for reliable and valid data to decide three questions.

> Do students with specially trained teachers achieve more than students in the regular writing program?
>
> Do the specially trained teachers have a positive spin-off effect on the achievement of students in their building whose teachers received no direct special training?
>
> Do students in the regular program achieve measurable growth in writing skill in one year?

The pretest was administered during early September, 1980, with the posttest data collected during late May, 1981. The tests were hand scored, data keypunched, and the data was analyzed. The statistical test chosen was analysis of covariance. The pretest served as the covariate while the students' grade level and the research group to which they were assigned served as classification variables.

This procedure has three advantages over other possible sta-

tistical treatments of the data. First, it statistically adjusts for all pretest differences so that in final data interpretation it can be assumed that all comparison groups were equal in the beginning. This adjustment assures that any significant differences on the posttest can be assumed to be related to happenings during the treatment period rather than any preexisting conditions.

Secondly, since complete pretest and posttest data are available for all students in the study, the knowledge gained by knowing students' pretest scores is used to increase the power of the statistical procedure. Third, it bypasses the statistical problems with the unreliability of difference scores that exists when using the differences between posttest and pretest scores (repeated measures t-test procedure) to analyze the data. The designs can be schematically represented.

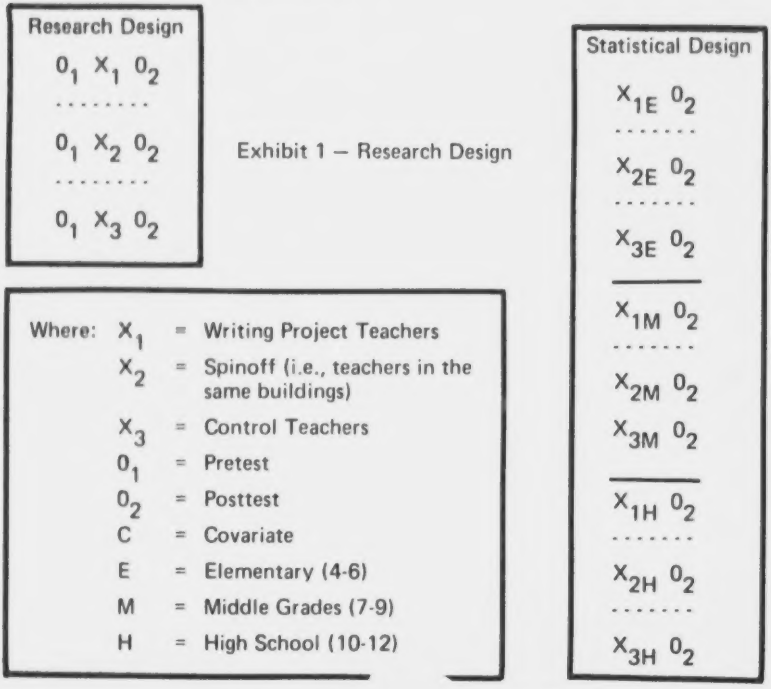


Exhibit 1 — Research Design

Chapter IV

Sampling Procedures

The experimental group consisted of all students taught by the 12 trained teachers. Two types of control groups were established: 1) students in schools having Writing Project trained teachers and 2) students from randomly selected schools not having Writing Project trained teachers. By the use of these three distinct groups, it was possible to determine the direct impact of the Writing Project on students, and, additionally, to measure the impact of trained teachers on writing instruction in their building through their collegial influence.

The entire selected sample of students whose teachers were in the experimental group or one of the two control groups were asked to write three short pretest essays and three short posttest essays. Only those students who completed all six essays were included in the data producing sample. The total numbers of students included in the final analyses are given in Table 2.

LEVEL	GROUP			TOTAL
	Writing	Spinoff	Control	
Elementary	39	64	231	334
Middle	251	158	147	556
Secondary	126	508	45	679
Total	416	730	423	1569

Table 2 – Data Producing Samples

The first pretest sample had 3933 essays, the second, 3867; and the third, 3792. For the posttest, the first sample had 2062 essays; the second, 2026; and the third, 2002. These numbers of samples were very large. However, only about 52 percent of the selected students had all three pretests and all three posttests. For the most part, this loss represented control teachers who chose not to provide the posttest data.

Ideally, 100 percent of the students in the sample should be data producing. However, within the realities of school-based research, generally teachers cannot be forced to provide requested information. In this type of research in which some teachers are being asked to voluntarily interrupt their teaching schedules to administer tests for a program in which they were not involved, a high mortality rate can be anticipated. By initially oversampling, a sufficiently large data producing sample was obtained for statistical analysis.

Were the mortality rate to cause a problem with the evaluation results, it could be expected to cause a Type II statistical error, or tend to show no differences when growth actually occurred.

It could be reasoned that those teachers who followed through with their commitment to provide posttest results could be expected, if anything, to be the control teachers most positive about their students' writing behavior. If this condition were to exist, it would tend to enhance the scores of the no treatment control group, making it more difficult, not less, to show positive treatment results.

The remaining students dropped from the data producing sample were those whose identification numbers were not discernable, another 12% of the initial selected sample.

It should be noted that all essays written by students were graded and recorded. However, only the data producing samples were statistically analyzed.

Chapter V

Data Collection

Several issues must be addressed in determining data collection procedures. Of prime importance is to maintain the integrity of the data in terms of objectivity of the graders and noncontaminated responses from students. These issues can be resolved by careful attention to the type of topic selected, the directions given to students, and the type of identification system used.

The topics selected should represent the types of writing normally expected of students and should provide sufficient flexibility for teachers to perceive the validity of the requested task. Thus, it would undoubtedly bias the results to key topics directly to the new curriculum or to provide such detailed prompts that only some students could recognize the task.

With holistic scoring, topics may be grade specific, or the same topics may apply to several different grade levels. Even among writing educators, there seems to be little consensus about the type of topics assigned. However, several factors pointed toward the use of a single topic. One set of scoring standards developed for assessing elementary through high school level essays allowed analysis of the interaction between writing treatment and grade level. The topics selected were sufficiently general so that all

levels of students could identify with the tasks. Also, both trained and control teachers could attest to the validity of the task. The topics tended to elicit the same type of expressive/transactional writing. Finally, it was decided that the topics should be singular, with each specifying one issue, concern or direction for response. A topic had to be broad enough so that all students, regardless of background, would have an opportunity to formulate a response.

Consistent with these criteria, the following topics were selected:

1. **If I were elected President, the first thing I would do is . . . . . .**

2. **My worst (or best) day at school was . . . . . .**

3. **The best gift I ever received was . . . . . .**

Directions

Once the topic is selected, the directions for administration must be prepared. The specificity of these directions depends upon the purpose for the assessment. If the assignment is to be criterion scored for grade level competency, the specific, detailed directions establishing every aspect of the process would be suitable. On the other hand, if the assessment is to comparatively reflect the "typical" behavior of students in different treatment programs, then the directions must be open enough to reflect the modalities of those programs. To force students into the framework of one program when they have undergone the alternate treatment would force a bias in the results. Thus, in the second instance, the directions must be sufficient to enable all teachers to understand the task and to administer the

Writing Samples:

DUE DATE:	September 30 if convenient with your schedule.
INSTRUCTIONS:	Teachers, please ask your students to write approximately one paragraph on each of the following: (Sentence stem) If I were elected president, the first thing I would do is . . . . . . (Sentence stem) My worst (or best) day at school was . . . . . . (Sentence stem) The best gift I ever received was . . . . . .
IDENTIFICATION:	Each of these writing samples should be written on separate sheets of paper.
ON THE BACK:	Students should provide the following information: Student Name Teacher Name Grade School Name Date

Exhibit 2 – Writing Directions

tests under similar circumstances, yet maintaining integrity within the classroom setting.

Within this framework, all treatment and control teachers were sent the same direction sheet specifying that on three separate sheets, students were to write on each of the three essay topics, allowing 10 minutes per topic. Students were to put their names, the teacher's name, school and grade level at the top of each sheet. In addition, teachers were encouraged to obtain makeup essays from absent students within a few days of the test administration. The directions were simple with no writing "prompts" suggested.

Contrary to some recommendations (Popham, 1978), the same questions were administered to all classes, both pre and post. This procedure maximized the consistency in scoring and analysis. To have utilized alternate topics or to have changed topics from fall to spring testing would have clouded the results by adding additional factors for potential research validity problems. It would have been difficult to screen out such factors to obtain a clear look at the treatment impact.

Chapter VI

Data Preparation

The six essays contributed by each student were combined into one group for scoring. This procedure eliminated any concern for potential scoring bias since readers would have no knowledge about the pretest or posttest nature of the papers nor be able to identify treatment or control papers. Yet, the essays were more easily scored by topic so that readers would not be constantly shifting their frame of reference from one topic to another. The pre and post essay responses were randomized within the three essay topics.

The randomization process called for assigning each paper a random number. The papers were then sequentially ordered by the random numbers. This process was the most cumbersome of the total evaluation process. To expedite the procedure, 30,000 labels were printed using the UNIFORM function provided by Statistical Analysis System (SAS) for generating pseudo-random numbers. The labels were printed in two sets, three across and 500 deep. The information on the labels included:

- a single digit number from 1-3 to identify the essay type
- a single digit number from 0-9, with 0-4 indicating a pretest and 5-9, posttest

23

- a four digit student number running from 0001 to 5000

- a seven digit random number with a possible range of 0000001 to 9999999.

The essay type identifier was sequential from 1 to 3 going across the three labels. The pretest-posttest identifier varied from 0 to 4 for the first set of labels and 5 to 9 for the second set. The student numbers were in serial order going down the labels with the numbers repeated going across the three labels. The student numbers were repeated on the second set of labels. The random numbers were different for all 30,000 labels.

7003456-3 00201 First	3100218-4 00201 Second	83627167-1 00201 Third
75321076-5 00202 First	63452178-9 00202 Second	1789671-6 00202 Third
02221702-3 00203 First	21135142-4 00203 Second	53765516-0 00203 Third

Exhibit 3 – Mock-up of Statistical Analysis System Generated Labels for Randomization

For each pretest essay, the student, teacher and other identification were transferred to coding sheets for keypunching. The label was placed over the identifying information on the face of the essay.

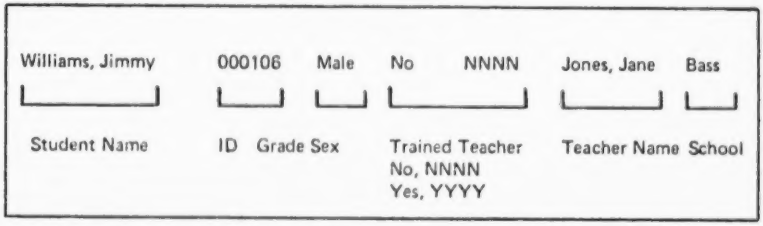


Exhibit 4 – Sample ID Data Card

After all pretest essays were prepared and the coded data keypunched, an alphabetical listing of the students and their identification numbers was generated for use in preparing the posttest essays. During late May, the preparation process was repeated for the posttests. This time, however, there was no need to code the student identification information since that had been done from the pretest. It was required, though, to match the student names on the essays with those on the printout to identify the student numbers before labeling.

After the labeling was completed, all essays were sorted by topics by the random numbers. This provided a thorough mix of pretest and posttest and treatment and control papers within each reading sequence.

This total preparation process, excluding keypunching, took about 200 hours, or about 41 seconds per essay.

Certainly one must, in the process of preparing so many papers for grading, keep in mind that the end result, maintaining the integrity of the data, justifies the time and expense required.

Chapter VII

Scoring

After the writing samples were completed, collected and prepared, the scoring process was begun. A scoring rubric was established for processing the papers, and a procedure for data reduction was determined. The scoring process consisted of four different stages or functions:

- anchoring the papers to determine the rubric
- training the readers
- actually scoring the essays
- recording the scores for keypunching.

ANCHORING - In order to examine the writing treatment effect across grade levels, a single rubric of six points was established for scoring the essays. Several weeks before the scoring was to be done, a core group consisting of the project staff and outside consultants selected from the set to be graded the prototypes, or anchor papers, that they felt were representative of each scoring point on the six-point scale. Sorting, rating and discussions of papers continued until there was over 80 percent consistency reached on the prototypes by the individuals developing the

rubric. By examining the classified papers describers for the rubric were set. The anchor papers were then reproduced to use for training the scorers.

POINT SCORE	CHARACTERISTICS FOR 3-12 WRITING ASSESSMENT
6	Central focus and several effective reasons or specific details Fully developed paper with clear sense of order Fluent in words and/or ideas Originality of thought or expression Generally clear mechanics
5	Central idea and specific facts, details, or reasons Consistent development Generally clear mechanics – errors do not interfere with overall effectiveness Less insightful, imaginative, concrete, or developed than "6"
4	Several clear ideas or specifics relevant to topic Evidence of fluency but not of unified development May be overly general or trite May have simple sentence structure or vocabulary Mechanical errors do not affect readability
3	At least one idea with some supporting material or several unrelated ideas Less fluent, developed, or detailed than "4" Sentences, vocabulary and thought may be simplistic Mechanical errors do not seriously affect readability or more serious mechanical problems accompany above-average content and fluency
2	At least one relevant idea Fluency and thought are minimal May have many mechanical errors, but paper is readable
1	May be almost devoid of content and/or totally inadequate in development May be unreadable due to spelling, handwriting, or other mechanical problems

Concrete details or originality will generally raise a paper 1 point

Exhibit 5 – Scoring Rubric

When the two scores for an essay were more than one point apart, the essay was placed in a re-score stack. These essays were periodically read by either of two lead readers to determine a third score that could agree with either of the first two readers or be a compromise between them. Very few of the essays had to be re-scored.

The detailed records for each student consisting of each scorer's marks and the tie-breaking scores on those papers that required it produced a more complete data picture for analysis. In some holistic scoring endeavors, only the average of the two scorer's marks is retained, thus losing some degree of data.

An early administrative task was the determination of the number of readers needed and an estimation of the total amount of scoring time required. In preliminary work, it was estimated that each paper would take about 45 seconds or 0.75 minutes to read for each of two readers, inclusive of any third reading caused by a discrepancy of more than one point on the rubric scale. In addition, it was estimated that six hours per day would be the maximum time for reading that a person could stay on task. Finally, about 15 minutes of each hour would be set aside for anchoring and break, leaving 45 minutes for scoring. Thus, the Total Reader Days (TRD) needed to score essays would approximate:

$$TRD = \frac{\text{Total Papers} \times 0.75(\text{Min./Reading}) \times 2(\text{Readings})}{6(\text{Hours}) \times 45(\text{Minutes})} + .5$$

The total number of essay to be scored was 17,682, yielding an estimated Total Reader Days of

$$TRD = \frac{17{,}682 \times 0.75 \times 2}{6 \times 45} + .5 = 98.73 \text{ or } 99.$$

This translates into $\frac{99}{5}$ =19.8 teachers for one week or $\frac{99}{10}$ = 9.9 or 10 teachers for two weeks.

The actual number of readers recruited was 12 for an esti= mated 8.25 days. The actual scoring process took eight days.

TRAINING - Training for scoring the essays consisted of two primary activities. The readers were first introduced to the holistic scoring concept, including how it is done, and why it rather than more traditional procedures was to be used. The second stage of training was the introduction of the particular rubric they were to use. The scorers practiced scoring the anchor papers until 80 percent concensus was reached.

The training for the readers took the first half day. Of the 12 readers who were recruited from a nearby school system, 11 reached agreement quite rapidly. All twelve reached agreement in succeeding scoring rounds.

SCORING - As the actual scoring began, each paper was read by two readers. A stack of 50-75 papers was placed at the center of a table. Readers took papers sequentially from the stack, read the papers, marked the scores (1-6) on the back and put the scored papers in a second stack. As a table completed a round of first readings, they were given a stack of 50-60 papers from another table for a second reading. This time the score was placed on the front of the essay

RECORDING - After both first and second readings were completed for a set of papers, the completed stack was transferred to the recorders. The identification number, both scores, pretest-posttest code, and essay type from each essay was coded on data entry sheets for enventual keypunching.

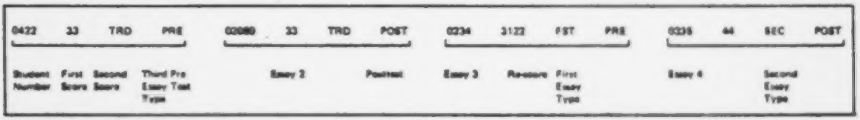


Exhibit 6 - Sample Score Data Card

Chapter VIII

Data Analysis

After all the scoring was complete and the data keypunched, the results were analyzed using analysis of covariance for differences in posttest scores, covaring pretest scores. Four 3 (grade levels) by 3 (groups) factorial designs were used in the analysis. The results for each essay and the total score for each student's complete set of essays were run separately.

All analyses showed significant groups and grade differences (see Table 3). There were no interaction effects. The adjusted posttest means are presented in Table 4 and non-adjusted pretest and posttest means are given in Table 5. The least squares probabilities were determined for the planned comparisons among grade levels and among treatment groups.

It can be seen in Table 4 that the students of the trained teachers obtained the highest mean scores; the students of the non-trained teachers in the same buildings, the second highest mean scores; and the students of the control teachers, the lowest mean scores. All these differences were significant. This clearly documents the program's impact.

Students of teachers that went through the writers workshop

ESSAY	SOURCE OF VARIANCE	SS	DF	MS	F
FIRST	Between Groups	280.54	8		
	Treatment	172.75	2	86.38	30.64**
	Grade	97.34	2	48.67	17.26**
	Grade by Treatment	10.45	4	2.61	0.48
	Within	4395.25	1559	2.82	
SECOND	Between Groups	294.44	8		
	Treatment	231.41	2	115.70	39.20**
	Grade	54.34	2	27.17	9.20**
	Grade by Treatment	8.69	4	2.17	0.74
	Within	4602.16	1559	2.95	
THIRD	Between Groups	307.95	8		
	Treatment	220.61	2	110.31	38.16**
	Grade	63.35	2	31.68	10.96**
	Grade by Treatment	23.99	4	6.00	2.08
	Within	4506.62	1559	2.89	
TOTAL	Between Groups	2155.50	8		
	Treatment	1801.44	2	900.72	60.71**
	Grade	271.38	2	135.69	9.15**
	Grade by Treatment	82.68	4	20.67	1.39
	Within	23131.17	1559	14.84	

Table 3 – Factorial Analyses of Covariance

ESSAY	LEVEL	TREATMENT GROUP Control	Spinoff	Trained	TOTAL*
FIRST#	Elementary	5.61	5.84	6.96	6.14
	Middle	5.65	6.22	6.91	6.26
	Senior High	6.68	6.73	7.44	6.95
	TOTAL*	5.98	6.27	7.11	6.47
SECOND#	Elementary	5.46	5.91	7.04	6.14
	Middle	5.78	6.46	6.98	6.41
	Senior High	6.12	6.91	7.41	6.82
	TOTAL*	5.79	6.43	7.14	6.54
THIRD#	Elementary	5.15	6.36	6.66	6.06
	Middle	5.65	6.25	7.11	6.34
	Senior High	6.30	6.84	7.22	6.79
	TOTAL*	5.70	6.48	7.00	6.45
TOTAL+	Elementary	16.81	18.21	20.96	18.66
	Middle	17.09	19.04	21.45	19.19
	Senior High	18.86	20.08	21.61	20.19
	TOTAL*	17.59	19.11	21.34	19.47

* The total level or group means underlined (or with bar) are not significantly different, = .05.
\# Based on total of 12 points.
\+ Based on total of 36 points.

Table 4 – Adjusted Posttest Treatment and Grade Level Essay Means

ESSAY	LEVEL	PRETEST			POSTTEST		
		Control	Spinoff	Trained	Control	Spinoff	Trained
FIRST	Elementary	4.04	5.19	4.64	5.10	5.75	6.67
	Middle	5.37	5.27	4.35	5.62	6.16	6.65
	Senior High	5.80	6.52	6.62	6.80	7.12	7.86
SECOND	Elementary	4.26	5.14	4.74	5.09	5.83	6.82
	Middle	5.40	5.08	4.61	5.79	6.35	6.72
	Senior High	5.87	6.18	6.33	6.29	7.18	7.73
THIRD	Elementary	4.41	5.50	5.07	4.71	6.36	6.49
	Middle	5.64	5.30	4.56	5.70	6.16	6.74
	Senior High	6.24	6.28	6.35	6.58	7.14	7.55
TOTAL	Elementary	12.71	15.83	14.46	14.90	17.94	19.97
	Middle	16.41	15.66	13.53	17.12	18.68	19.98
	Senior High	17.91	18.98	19.30	19.67	21.44	23.14

Table 5 – Non-Adjusted Pretest and Posttest Means

learned how to write better than students of other teachers. It is particularly interesting to note that the students of the "spin-off" teachers, those teachers who did not attend the workshop, but were in buildings with teachers who did attend, were significantly better writers than students of the control teachers, but significantly poorer writers than students of the trained teachers. This result suggests that the trained teachers successfully communicated the writers program to their colleagues. Teacher interactions can be effective in transmitting the writer program philosophy and teaching strategies.

The grade level results reflected a similar story. Overall, the high school students ranked first with the highest mean scores, followed by the junior high students and, at the bottom, the elementary students. Of note, though, while the high school students were significantly higher than the junior high and elementary students, there was no significant difference between these latter two levels. It seems that overall, there is no real difference between the writing skills of elementary and junior high students, and a small but significant difference between high school students and students in lower grades. This result may reflect a general lack of "writing" instruction, apart from practice with spelling, grammar and punctuation, prior to high school.

The graphic representation of the results (see Figure 1) illustrates further the impact of the writers program. There seems to be greater differences among groups, particularly the students of the control and trained teachers, than among grade levels. In addition, while the major impact on writing for the control students seemed to be at the high school level, the major impact for the writers project students seemed to be at the elementary and junior high levels. In the absence of an interaction effect, this latter supposition needs additional study. If this trend of

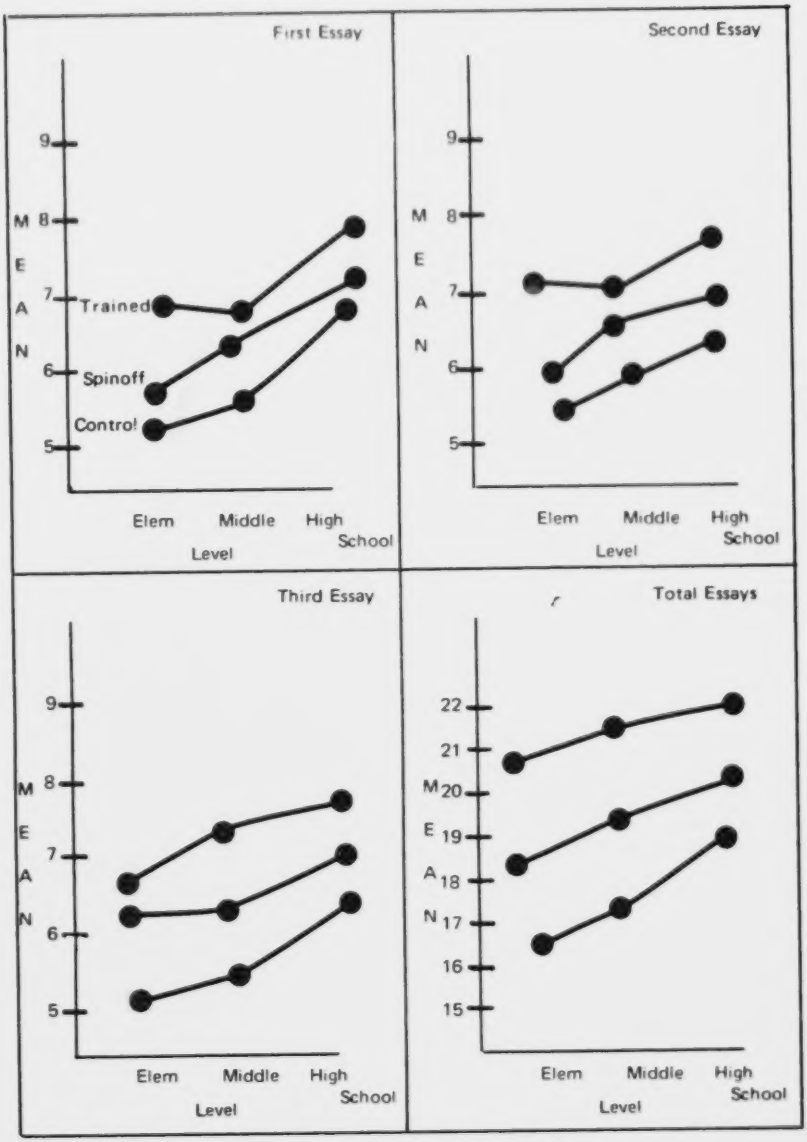


Figure 1 — Adjust Posttest Means on the Three Essays and Total Score

high elementary and junior high achievement continues, it may be necessary in the future to reformulate goals and strategies for high school writing instruction.

Of final note, the reliability of scoring was consistently high using as estimates the percent of exact agreement and the percent agreement within one point. The consistency was checked hourly, at each re-anchoring. For exact agreement, the reliabilities ranged from a low of 44 percent to a high of 92 percent for the hourly checks, with a median of 54 percent; for agreement within one point, the range was from 80 percent to 100 percent, with a median of 96 percent.

For the overall scoring, the essays were scored within one point for a general reliability estimate of 98.14 percent and with exact agreement at 80.24 percent.

Chapter IX

Conclusions

Relating the results to the two purposes for the evaluation, a) determining the impact of a new approach to writing instruction and b) determining the amount of writing growth typically occurring, it can be concluded that the new approach to writing has a significant impact on students at all levels. Further, student growth in writing does occur during a one year period. The levels of writing demonstrated by the end of the first year are described in Table 6.

The increase in student achievement that was documented is perhaps the most supportable outcome of the study. In addition, there are other conclusions evident that tended to confirm certain principles:

Principle 1: **Pretest-posttest, experimental-control group research designs can be effectively used to assess the effects of "new" programs on students.**

Principle 2: **A meaningful rubric for holistic scoring can be developed for students in grades 3-12.**

Level Students	Students of writing project teachers	Students of other teachers in buildings with project teachers	Students of control teachers
Elementary Level	Generally one idea with supportive material, or several unrelated ideas; some evidence of fluency; sentences, vocabulary thought generally simplistic; mechanical errors do not seriously affect readability.	Generally one idea with supportive material, or several unrelated ideas; little evidence of fluency; sentences, vocabulary and thought generally simplistic; mechanical errors do not seriously affect readability.	At least one relevant idea; fluency minimal; sentences, vocabulary and thought generally simplistic; may have many mechanical errors, but paper is readable.
Middle Level	Generally one or more clear ideas or specifics relative to topic; some evidence of fluency; often overly general; sentences, vocabulary and thought may be simplistic; mechanical errors do not affect readability.	Generally one idea with supportive material, or several unrelated ideas; little evidence of fluency; sentences, vocabulary and thought generally simplistic; mechanical errors do not seriously affect readability.	At least one relevant idea; fluency minimal; sentences, vocabulary and thought generally simplistic; may have many mechanical errors, but paper is readable.
Senior High Level	Generally one or more clear ideas or specifics relative to topic; some evidence of fluency; often overly general; sentences, vocabulary and thought may be simplistic; mechanical errors do not affect readability.	Generally one idea with supportive material, or several unrelated ideas; some evidence of fluency; sentences, vocabulary thought generally simplistic; mechanical errors do not seriously affect readability.	Generally one idea with supportive material, or several unrelated ideas; little evidence of fluency; sentences, vocabulary and thought generally simplistic; mechanical errors do not seriously affect readability.

Table 6 – Descriptive Levels of Student Achievement for the Writers Project

Principle 3: **Holistic scoring can be reliably used to assess large numbers of essays.**

Principle 4: **Using the Bay Area Writing Process for teacher training, with teacher support systems throughout the school year, will result in improved student writing achievement.**

Principle 5: **After only one year, project students in the elementary grades are writing at a level higher than the current high school students. This could suggest a need for new thinking in developing writing curriculum for future high school students.**

Principle 6: **Staff development for teachers can result in dramatic changes in student behavior.**

Principle 7: **The program development and adoption process can result in the successful diffusion of programs that can result in the improved basic skill achievements of students.**

As the program implementors observed the process of instituting a new program, they concluded that the most significant change resulting from the new emphasis on writing is that teachers have begun to see themselves as writers. A sense of wonder at their own talent that many did not previously recognize in themselves fortifies a deeply felt commitment to the teaching of writing. Through the utilization of the more effective teaching strategies, this same phenomenon occurs within students. They, too, begin to see themselves as writers.